ACADEMIC PRESS
Journal of Computer and System Sciences 65 (2002) 612–625
JOURNAL OF COMPUTER AND SYSTEM SCIENCES
www.academicpress.com

A linear lower bound on the unbounded error probabilistic communication complexity

Jürgen Forster

Lehrstuhl Mathematik & Informatik, Fakultät für Mathematik, Ruhr-Universität Bochum, 44780 Bochum, Germany

Received 23 July 2001; received in revised form 6 May 2002

Abstract

The main mathematical result of this paper may be stated as follows: Given a matrix $M \in \{-1,1\}^{n\times n}$ and any matrix $\tilde{M} \in \mathbb{R}^{n\times n}$ such that $\operatorname{sign}(\tilde{M}_{i,j}) = M_{i,j}$ for all i,j, then $\operatorname{rank}(\tilde{M}) \geqslant n/\|M\|$. Here $\|M\|$ denotes the spectral norm of the matrix M.

This implies a general lower bound on the complexity of unbounded error probabilistic communication protocols. As a simple consequence, we obtain the first linear lower bound on the complexity of unbounded error probabilistic communication protocols for the functions defined by Hadamard matrices. This solves a long-standing open problem stated by Paturi and Simon (J. Comput. System Sci. 33 (1986) 106).

We also give an upper bound on the margin of any embedding of a concept class in half spaces. Such bounds are of interest to problems in learning theory.

Keywords: Lower bounds; Probabilistic communication complexity; Hadamard matrix; Spectral norm

0. Introduction

Lower bounds on the complexity of communication protocols have applications in several areas such as circuit complexity, data structures and VLSI. For a list of references, see Kushilevitz and Nisan [13]. In this paper, we prove new and improved lower bounds on the communication complexity of distributed functions. In particular we show that the unbounded error probabilistic communication complexity of the functions defined by Hadamard matrices is linear. This solves a long-standing open problem stated by Paturi and Simon [14] and Krause [12].

In this paper, a probabilistic communication protocol is a probabilistic algorithm for two processors P_0 and P_1 that computes a distributed function $f : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$. Both processors have unbounded computational power. Processor P_0 sees only the first part, x, and P_1 sees only the last part, y, of the input $(x,y) \in \{0,1\}^n \times \{0,1\}^n$. Obviously, there has to be some communication between the two processors to calculate $f(x,y) \in \{0,1\}$. The processors can communicate by exchanging messages $b \in \{0,1\}^*$. The computation takes place in *rounds*. In each

E-mail address: forster@lmi.ruhr-uni-bochum.de (J. Forster).

0022-0000/02/$ - see front matter
PII: S0022-0000(02)00019-3

round, one of the processors is active, in odd rounds it is P_0 and in even rounds, it is P_1. The active processor probabilistically (depending on the part of the input it knows and on the past messages) chooses a message according to the communication protocol. In the final round, the active processor probabilistically chooses the result of the computation.

We say that a protocol computes the distributed function $f : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ with unbounded error if for all inputs $(x,y) \in \{0,1\}^n \times \{0,1\}^n$ the correct output is calculated with probability greater than $\frac{1}{2}$. The complexity of a communication protocol is $\lceil \log_2 N \rceil$, where N is the number of distinct message sequences that can occur in computations that follow the protocol. The communication complexity C_f of a distributed function $f : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ is the smallest complexity that a communication protocol for f can have.

It is known that the unbounded error probabilistic communication complexity for almost all functions $f : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ is linear in n (see [1,14]). This was shown by counting arguments that do not give lower bounds on the communication complexity of explicit functions. Paturi and Simon [14] give an example of a distributed function with logarithmic communication complexity. They conjecture that the functions defined by Hadamard matrices have linear probabilistic communication complexity. We prove this conjecture in Corollary 2.2.

Our techniques can also be applied to another class of problems. Recently, there has been a lot of interest in maximal margin classifiers. Learning algorithms that calculate a hyperplane that separates positive and negative instances of a sample with the largest margin and use this hyperplane to classify new instances have shown excellent empirical performance (see [4]). Often the instances are mapped (implicitly when a kernel function is used) to some possibly high-dimensional space before the hyperplane with maximal margin is calculated. If the norms of the instances are bounded and a hyperplane with large margin can be found, a bound on the VC-dimension can be applied [4,16, Theorem 4.16]. A small VC-dimension means that a concept class can be learned with a small sample size [3,11,17] [3,11,17, Theorem 3.3].

The success of maximal margin classifiers raises the question of what concept classes can be embedded in half spaces with a large margin. For every concept class, there is a trivial embedding in half spaces. Ben-David et al. [2] show that most concept classes (even of small VC-dimension) cannot be embedded with a margin that is much larger than the trivial margin. They use counting arguments that do not give an upper bound on the margin for an explicit concept class. Vapnik [16] also showed an upper bound on the margin that is in terms of the VC-dimension. We prove a general upper bound on the margin that is much stronger than Vapnik's in the case of concept classes defined by Hadamard matrices. We show that for these only the trivial margin can be achieved.

Both a distributed function $f : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ and a concept class $\mathscr{C}$ over an instance space X can be represented by a matrix with entries ± 1. For a distributed function f we can use the matrix $\mathrm{M}_f := (2f(x,y)-1)_{x,y \in \{0,1\}^n}$, and for a concept class $\mathscr{C}$ over the instance space X we can use the matrix $\mathrm{M}_{X,\mathscr{C}} \in \{-1,1\}^{X \times \mathscr{C}}$ for which the entry $(\mathrm{M}_{X,\mathscr{C}})_{x,c}$ is 1 if $x \in c$ and -1 otherwise.

Krause [12] shows that the complexity of any probabilistic communication protocol which computes a function $f : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ with error probability bounded by $\frac{1}{2} - \frac{1}{s}$ is at least $\frac{1}{4}(n - \log_2 \|\mathrm{M}_f\| - \log_2 \sqrt{s} - 2)$ for all $s \in \mathbb{N}$. Here $\|\mathrm{M}_f\|$ is the spectral norm of the matrix M_f. We improve on this result by showing that the assumption that the error of the protocol is bounded is not really needed: In Section 2, we prove a lower bound of $n - \log_2 \|\mathrm{M}_f\|$ on the unbounded error communication complexity. A result from matrix theory needed in the proof is given in Section 4. In Section 3, we show that any concept class $(X, \mathscr{C})$ can only be

embedded in homogeneous half spaces with margin at most $\|M_{X,\mathscr{C}}\|/\sqrt{|X|\,|\mathscr{C}|}$. In the next section, we fix some notation for the rest of the paper.

After a preliminary version of this paper [5] was published, some of the techniques and results presented here have been improved and strengthened.

It is shown in [6] that the lower bound on the unbounded error probabilistic communication complexity of a distributed function given in Section 2 is reasonably good for almost all distributed functions: It is linear in n for the vast majority of distributed functions.

The lower bound on the dimension of arrangements of half spaces given in Theorem 2.2 is used in [6] to show lower bounds on the size of depth-2 threshold circuits that compute Hadamard matrices: If the top gate of the circuit is a linear threshold gate with *unrestricted* weights and if there are s linear threshold gates on the bottom level with integer weights of absolute value at most W, then $s = \Omega(2^{n/2}/(nW))$. If the top gate of the circuit is a linear threshold gate with *unrestricted* weights and there are s gates on the bottom level that compute symmetric functions, then $s = \Omega(2^{n/2}/n)$. Also, Theorem 2.2 is used to give an example of a function which can only be computed by probabilistic ordered binary decision diagrams (OBDDs) of exponential size.

Furthermore, the lower bound of Theorem 2.2 is generalized in [6]. An example of a class of matrices is given for which Theorem 2.2 fails, but for which the generalization of Theorem 2.2 still gives a strong lower bound on the dimension. It is shown that the dimension k of any arrangement of half spaces realizing a matrix $M \in \{-1,1\}^{X\times Y}$ is at least $k \geqslant \sqrt{|X|\,|Y|}/\|\tilde{M}\|$ for any matrix $\tilde{M} \in \mathbb{R}^{X\times Y}$ for which the entries $\tilde{M}_{x,y}$ have absolute value at least 1 and which has the same sign pattern as M (i.e. $\operatorname{sign}(\tilde{M}_{x,y}) = M_{x,y}$ for all $x \in X$, $y \in Y$). This bound is more general than Theorem 2.2 because the absolute values of the entries of $\tilde{M}$ do not have to be equal.

A similar generalization of Theorem 3.1 to matrices with arbitrary nonzero entries is given in [7]. There it is shown that the margin γ of any embedding of a matrix $M \in \{-1,1\}^{X\times Y}$ is at most

$$\gamma \leqslant \frac{\sqrt{|X|}\|\tilde{M}\|}{\sqrt{\sum_{y\in Y}(\sum_{x\in X}|\tilde{M}_{x,y}|)^2}}$$

for any matrix $\tilde{M} \in \mathbb{R}^{X\times Y}$ with the same sign pattern as M. This result is used to show that the optimal margin for embedding the matrix $M \in \{-1,1\}^{n\times n}$ with $M_{i,j} = 1$ iff $i \geqslant j$ is of the order $\frac{\pi}{2\ln n} + \Theta(\frac{1}{(\ln n)^2})$.

1. Preliminaries from matrix theory

For a finite set X, $\mathbb{R}^X$ is the vector space of real-valued functions (“vectors”) on X. The vectors $v \in \mathbb{R}^X$ are column vectors, and the Euclidean norm of v is $\|v\| := \sqrt{\sum_{x\in X} v_x^2}$. For two finite sets X, Y we write $\mathbb{R}^{X\times Y}$ for the set of matrices with rows indexed by the elements of X and columns indexed by the elements of Y. As usual, we write $\mathbb{R}^n := \mathbb{R}^{\{1,\ldots,n\}}$ and $\mathbb{R}^{m\times n} := \mathbb{R}^{\{1,\ldots,m\}\times\{1,\ldots,n\}}$. The identity matrix in $\mathbb{R}^{X\times X}$ is denoted by I_X, the identity matrix in $\mathbb{R}^{n\times n}$ by I_n. The transpose of a matrix $A \in \mathbb{R}^{X\times Y}$ is $A^\top \in \mathbb{R}^{Y\times X}$. We define $e_x \in \mathbb{R}^X$, $x \in X$, to be the canonical unit vectors satisfying

$$(e_x)_y = \begin{cases} 1, & x = y, \\ 0, & x \neq y, \end{cases}$$

for $x, y \in X$. The $(k-1)$-dimensional sphere, i.e. the set $\{v \in \mathbb{R}^k \mid \|v\| = 1\}$, is denoted by S^{k-1}.

For a linear function $A : \mathbb{R}^X \to \mathbb{R}^Y$, the null space is

$$\text{null}(A) = \{v \in \mathbb{R}^X \mid Av = 0\},$$

and the range of A is

$$\text{range}(A) = \{Av \mid v \in \mathbb{R}^X\}.$$

The spectral norm of a matrix $A \in \mathbb{R}^{X \times Y}$ is

$$\|A\| = \sup_{\substack{v \in \mathbb{R}^X \\ \|v\| \leqslant 1}} \|Av\| = \max_{\substack{v \in \mathbb{R}^X \\ \|v\| \leqslant 1}} \|Av\|. \tag{1}$$

The supremum is attained because $\|Av\|$ is a continuous function of v and the unit ball $\{v \in \mathbb{R}^X \mid \|v\| \leqslant 1\}$ is compact. It is well known that for any matrix $A \in \mathbb{R}^{X \times Y}$: $\|A\|^2 = \|A^\top A\| = \|AA^\top\|$.

The trace of a square matrix $A \in \mathbb{R}^{X \times X}$ is $\text{trace}(A) = \sum_{x \in X} A_{x,x}$. A matrix $A \in \mathbb{R}^{X \times X}$ is called *symmetric* if $A^\top = A$. The *spectral theorem for symmetric matrices* [10, Theorem 2.5.6] states that every symmetric matrix is orthogonally diagonalizable. Equivalently, we can say that for every symmetric matrix $A \in \mathbb{R}^{X \times X}$ there is an orthonormal basis $d_1, \ldots, d_{|X|}$ of $\mathbb{R}^X$ consisting of eigenvectors of A and there are real numbers $\lambda_1, \ldots, \lambda_{|X|} \in \mathbb{R}$ (the eigenvalues of A) such that

$$A = \sum_{i=1}^{|X|} \lambda_i d_i d_i^\top.$$

In this case, the spectral norm $\|A\|$ is equal to the maximum of the absolute values of the eigenvalues of A. Also note that for any vector $v \in \mathbb{R}^X$ with norm $\|v\|^\top = 1$ the term $v^\top Av$ lies in the interval

$$v^\top Av \in \left[\min_{i=1}^{|X|} \lambda_i, \max_{i=1}^{|X|} \lambda_i\right] \tag{2}$$

because $v^\top Av = \sum_{i=1}^{|X|} \lambda_i \langle v, d_i \rangle^2$ is a convex combination of the eigenvalues (since $\sum_{i=1}^{|X|} \langle v, d_i \rangle^2 = \|v\|^2 = 1$).

A matrix $A \in \mathbb{R}^{X \times X}$ is said to be positive semidefinite if it is symmetric and $v^\top Av \geqslant 0$ for all $v \in \mathbb{R}^X$. For a finite set of vectors $u_x \in \mathbb{R}^k$, $x \in X$, the matrix $A := (\langle u_x, u_{\tilde{x}} \rangle)_{x,\tilde{x} \in X}$ is an example of a positive semidefinite matrix: Obviously, A is symmetric, and for an arbitrary vector $v \in \mathbb{R}^X$ we have

$$v^\top Av = \sum_{x,\tilde{x} \in X} v_x \langle u_x, u_{\tilde{x}} \rangle v_{\tilde{x}} = \left\langle \sum_{x \in X} v_x u_x, \sum_{\tilde{x} \in X} v_{\tilde{x}} u_{\tilde{x}} \right\rangle = \left\| \sum_{x \in X} v_x u_x \right\|^2 \geqslant 0.$$

For an arbitrary matrix $A \in \mathbb{R}^{X \times Y}$, the matrix $B := \|A\|^2 I_X - AA^\top$ is another example of a positive semidefinite matrix: Obviously, B is symmetric, and for an arbitrary vector $v \in \mathbb{R}^X$ we have $v^\top Bv = \|A\|^2 \|v\|^2 - \|Av\|^2$. This is nonnegative because $\|Av\| \leqslant \|A\| \, \|v\|$ by the definition (1) of the spectral norm.

We also need the following two results from matrix theory:

Theorem 1.1 (Fejer's Theorem [10, Corollary 7.5.4]). *A matrix $A \in \mathbb{R}^{X \times X}$ is positive semidefinite if and only if*

$$\sum_{x,\tilde{x} \in X} A_{x,\tilde{x}} B_{x,\tilde{x}} \geqslant 0$$

for all positive semidefinite matrices $B \in \mathbb{R}^{X \times X}$.

Proof. We only prove the part of the theorem that is used in this paper. Assume that $A, B \in \mathbb{R}^{X \times X}$ are positive semidefinite matrices. Because of the spectral theorem for symmetric matrices we can write $A = \sum_{i=1}^{|X|} \lambda_i d_i d_i^\top$ with an orthonormal basis $d_1, \ldots, d_{|X|}$ of $\mathbb{R}^X$. Because A is positive semidefinite we know that $\lambda_1, \ldots, \lambda_{|X|} \geqslant 0$. It follows that

$$\sum_{x,\tilde{x} \in X} A_{x,\tilde{x}} B_{x,\tilde{x}} = \sum_{i=1}^{|X|} \lambda_i \sum_{x,\tilde{x} \in X} (d_i)_x (d_i)_{\tilde{x}} B_{x,\tilde{x}} = \sum_{i=1}^{|X|} \lambda_i \underbrace{d_i^\top B d_i}_{\geqslant 0} \geqslant 0. \qquad \square$$

Bessel's Inequality (see [15]) states that for every set $d_1, \ldots, d_n \in \mathbb{R}^X$ of orthonormal vectors and every vector $v \in \mathbb{R}^X$:

$$\sum_{i=1}^{n} \langle v, d_i \rangle^2 \leqslant \|v\|^2. \tag{3}$$

In our bounds, the spectral norm of matrices $M \in \{-1, 1\}^{X \times Y}$ will appear. For a matrix of this form it is easy to see that $\|M\| = \sqrt{|Y|}$ if the rows of M are pairwise orthogonal and that $\|M\| = \sqrt{|X|}$ if the columns of M are pairwise orthogonal. Furthermore, $\operatorname{rank}(M) = 1$ iff $\|M\| = \sqrt{|X|\,|Y|}$.

Hadamard matrices $H_n \in \mathbb{R}^{2^n \times 2^n}$ are examples of matrices with pairwise orthogonal rows and pairwise orthogonal columns. They are recursively defined by

$$H_0 = 1, \quad H_{n+1} = \begin{pmatrix} H_n & H_n \\ H_n & -H_n \end{pmatrix}.$$

The signum function $\operatorname{sign}: \mathbb{R} \to \mathbb{R}$ is given by

$$\operatorname{sign}(x) = \begin{cases} 1, & x > 0, \\ 0, & x = 0, \\ -1, & x < 0. \end{cases}$$

2. A lower bound on the complexity of unbounded error probabilistic communication protocols

We say that a matrix $M \in \{-1, 1\}^{X \times Y}$ can be realized by an arrangement of homogeneous half spaces in $\mathbb{R}^k$ if there are vectors $u_x, v_y \in \mathbb{R}^k$ for $x \in X$, $y \in Y$ such that $\operatorname{sign} \langle u_x, v_y \rangle = M_{x,y}$ for all $x \in X$, $y \in Y$. A vector v_y (or analogously a vector u_x) can be interpreted as a normal vector of the boundary of the homogeneous half space $\{z \in \mathbb{R}^k \mid \langle z, v_y \rangle \geqslant 0\}$. Then $\operatorname{sign} \langle u_x, v_y \rangle = M_{x,y}$ means that the vector u_x lies in this half space iff $M_{x,y} = 1$.

The unbounded error probabilistic communication complexity C_f of a distributed function f is strongly related to the smallest dimension of any arrangement of homogeneous half spaces that realizes the matrix M_f:

Theorem 2.1 (Paturi and Simon [14, Theorem 2]). *Let $f : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ be a distributed function. If k is the smallest dimension of any arrangement of homogeneous half spaces that realizes M_f, then*

$$\lceil \log_2 k \rceil \leqslant C_f \leqslant \lceil \log_2 k \rceil + 1.$$

We state our main result now. It immediately implies a lower bound on communication complexities.

Theorem 2.2. *If a matrix $M \in \{-1,1\}^{X \times Y}$ can be realized by an arrangement of homogeneous half spaces in $\mathbb{R}^k$, then*

$$k \geqslant \frac{\sqrt{|X|\,|Y|}}{\|M\|}.$$

Proof. Assume that there is an arrangement of homogeneous half spaces in $\mathbb{R}^k$ realizing M, i.e. there are vectors $u_x, v_y \in \mathbb{R}^k$ such that $\operatorname{sign}\langle u_x, v_y \rangle = M_{x,y}$ for all $x \in X$, $y \in Y$. We can assume without loss of generality that $\operatorname{span}\{u_x \mid x \in X\} = \mathbb{R}^k$. This implies that $|X| \geqslant k$.

Furthermore, we can assume that any k of the vectors u_x, $x \in X$, are linearly independent. This can be seen as follows: We iteratively look at each vector u_x, $x \in X$, and modify it such that it does not lie in any linear span of $k-1$ of the other vectors $u_{\tilde{x}}$, $\tilde{x} \in X \setminus \{x\}$. This is possible because the signs of the scalar products $\langle u_x, v_y \rangle$ are not affected by small changes of the u_x, and the union of the linear spans

$$\bigcup_{x_1,\ldots,x_{k-1} \in X \setminus \{x\}} \operatorname{span}\{u_{x_1}, \ldots, u_{x_{k-1}}\}$$

is a set with Lebesgue measure zero.

Now we can apply a result from matrix theory that is proved in Section 4: Theorem 4.1 states that if $|X| \geqslant k$ and if $u_x \in \mathbb{R}^k$, $x \in X$, are vectors such that any k of these vectors are linearly independent, then there is a nonsingular linear transformation $A \in \mathbb{R}^{k \times k}$ such that

$$\sum_{x \in X} \tilde{u}_x \tilde{u}_x^\top = \frac{|X|}{k} I_k \tag{4}$$

for the vectors $\tilde{u}_x := \|Au_x\|^{-1} Au_x \in S^{k-1}$.

We also define transformed vectors $\tilde{v}_y := \|(A^\top)^{-1} v_y\|^{-1} (A^\top)^{-1} v_y \in S^{k-1}$, $y \in Y$. Then the vectors $\tilde{u}_x, \tilde{v}_y \in \mathbb{R}^k$ also realize the matrix M, because

$$\operatorname{sign}\langle \tilde{u}_x, \tilde{v}_y \rangle = \operatorname{sign}\langle Au_x, (A^\top)^{-1} v_y \rangle = \operatorname{sign}\langle u_x, v_y \rangle = M_{x,y}$$

for all $x \in X$, $y \in Y$. The new arrangement of half spaces has the advantage that the vectors $\tilde{u}_x$ are nicely balanced in the sense (4).

Now we have for all $y \in Y$ that

$$\sum_{x \in X} |\langle \tilde{u}_x, \tilde{v}_y \rangle| \overset{1 \geqslant |\langle \tilde{u}_x, \tilde{v}_y \rangle|}{\geqslant} \sum_{x \in X} \langle \tilde{u}_x, \tilde{v}_y \rangle^2 = \tilde{v}_y^\top \left(\sum_{x \in X} \tilde{u}_x \tilde{u}_x^\top \right) \tilde{v}_y \overset{(4)}{=} \frac{|X|}{k}. \tag{5}$$

Inequality (5) means that for all $y \in Y$ the absolute values of the scalar products $\langle \tilde{u}_x, \tilde{v}_y \rangle$ are on the average at least $\frac{1}{k}$. Therefore, the vectors $\tilde{u}_x$ cannot lie arbitrarily close to the homogeneous hyperplane with normal vector $\tilde{v}_y$.

Lemma 2.1 (which is proven below) gives a corresponding upper bound in terms of the spectral norm $\|M\|$ on the absolute values of the scalar products $\langle \tilde{u}_x, \tilde{v}_y \rangle$. It follows that

$$|Y| \left(\frac{|X|}{k} \right)^2 \overset{(5)}{\leqslant} \sum_{y \in Y} \left(\sum_{x \in X} |\langle \tilde{u}_x, \tilde{v}_y \rangle| \right)^2 \overset{\text{Lemma 2.1}}{\leqslant} |X|\,\|M\|^2. \qquad \square$$

From Theorems 2.1 and 2.2 (applied to the case $X = Y = \{0,1\}^n$, $M = \mathrm{M}_f$) we get the following corollary:

Corollary 2.1. *For every distributed function* $f : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$, *the communication complexity is at least*

$$C_f \geqslant n - \log_2 \|\mathrm{M}_f\|.$$

Recall that for a matrix $M \in \{-1,1\}^{X \times Y}$ with pairwise orthogonal columns, the spectral norm is $\|M\| = \sqrt{|X|}$. Now we can give an example of a distributed function that has linear communication complexity:

Corollary 2.2. *The communication complexity of the distributed function* $f : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ *for which* M_f *is an Hadamard matrix* H_n *is at least* $\frac{n}{2}$. *An Hadamard matrix* H_n *can only be realized by an arrangement of homogeneous half spaces in* $\mathbb{R}^k$ *if* $k \geqslant 2^{n/2}$.

We still have to prove the following lemma:

Lemma 2.1. *Let* $M \in \{-1,1\}^{X \times Y}$ *be a matrix and let* $u_x, v_y \in S^{k-1}$ *be vectors such that* $\operatorname{sign}\langle u_x, v_y \rangle = M_{x,y}$ *for all* $x \in X$, $y \in Y$. *Then*

$$\sum_{y \in Y} \left(\sum_{x \in X} |\langle u_x, v_y \rangle| \right)^2 \leqslant |X| \, \|M\|^2.$$

Proof. For every $y \in Y$ we have that

$$\begin{aligned} \sum_{x \in X} |\langle u_x, v_y \rangle| &= \sum_{x \in X} M_{x,y} \langle u_x, v_y \rangle \\ &= \left\langle \sum_{x \in X} M_{x,y} u_x, v_y \right\rangle \overset{\|v_y\|=1}{\leqslant} \left\| \sum_{x \in X} M_{x,y} u_x \right\|, \end{aligned} \tag{6}$$

where we used the Cauchy–Schwartz Inequality. We square inequality (6) and sum over $y \in Y$:

$$\begin{aligned} \sum_{y \in Y} \left(\sum_{x \in X} |\langle u_x, v_y \rangle| \right)^2 &\overset{(6)}{\leqslant} \sum_{y \in Y} \left\langle \sum_{x \in X} M_{x,y} u_x, \sum_{\tilde{x} \in X} M_{\tilde{x},y} u_{\tilde{x}} \right\rangle \\ &= \sum_{x,\tilde{x} \in X} (MM^\top)_{x,\tilde{x}} \langle u_x, u_{\tilde{x}} \rangle \overset{(*)}{\leqslant} \sum_{x,\tilde{x} \in X} (\|M\|^2 I_X)_{x,\tilde{x}} \langle u_x, u_{\tilde{x}} \rangle \\ &= \|M\|^2 \sum_{x \in X} \|u_x\|^2 = |X| \, \|M\|^2. \end{aligned}$$

For the proof of inequality $(*)$ we first note that $A := \|M\|^2 I_X - MM^\top$ and $B := (\langle u_x, u_{\tilde{x}} \rangle)_{x,\tilde{x} \in X}$ are positive semidefinite matrices (see Section 1). By Fejer's Theorem 1.1 we know that

$$0 \leqslant \sum_{x,\tilde{x} \in X} A_{x,\tilde{x}} B_{x,\tilde{x}} = \sum_{x,\tilde{x} \in X} (\|M\|^2 I_X)_{x,\tilde{x}} \langle u_x, u_{\tilde{x}} \rangle - \sum_{x,\tilde{x} \in X} (MM^\top)_{x,\tilde{x}} \langle u_x, u_{\tilde{x}} \rangle,$$

therefore $(*)$ holds. □

3. An upper bound on the margin of arrangements of half spaces

In this section, we are interested in the largest margin (not the smallest dimension) that an arrangement of half spaces that realizes a matrix $M \in \{-1,1\}^{X\times Y}$ can have. We say that the matrix $M \in \{-1,1\}^{X\times Y}$ can be realized by an arrangement of homogeneous half spaces with margin γ if there are vectors u_x, v_y for $x \in X$, $y \in Y$ that lie in the unit ball of $\mathbb{R}^k$ (where k can be arbitrarily large) such that $\operatorname{sign}\langle u_x, v_y\rangle = M_{x,y}$ and $|\langle u_x, v_y\rangle| \geqslant \gamma$ for all $x \in X$, $y \in Y$. If we interpret v_y as the normal vector of a homogeneous half space, then $\operatorname{sign}\langle u_x, v_y\rangle = M_{x,y}$ means that the vector u_x lies in the interior of this half space if and only if $M_{x,y} = 1$. The requirement $|\langle u_x, v_y\rangle| \geqslant \gamma$ means that the point u_x has distance at least γ from the boundary of the half space. Analogously, we can interpret the vectors u_x as normal vectors of half spaces and the vectors v_y as points. It is crucial that we require the vectors to lie in a unit ball (or that they are bounded) because otherwise we could increase the margin by simply stretching all vectors.

Note that it is not really a restriction to assume that the half spaces are homogeneous: Assume we have points $u_x \in \mathbb{R}^k$, $x \in X$, that lie in the unit ball and an arrangement of (not necessarily homogeneous) half spaces given by normal vectors $v_y \in S^{k-1}$ and thresholds $t_y \in [-1,1]$ such that

$$M = (\operatorname{sign}(\langle u_x, v_y\rangle - t_y))_{x\in X, y\in Y}.$$

Then the vectors

$$\tilde{u}_x := \frac{1}{\sqrt{2}}\begin{pmatrix} u_x \\ 1 \end{pmatrix}, \quad \tilde{v}_y := \frac{1}{\sqrt{2}}\begin{pmatrix} v_y \\ -t_y \end{pmatrix}$$

lie in the unit ball of $\mathbb{R}^{k+1}$, $M = (\operatorname{sign}(\langle \tilde{u}_x, \tilde{u}_y\rangle))_{x\in X, y\in Y}$, and the margin of the new arrangement is only by a factor of $\frac{1}{2}$ worse than the old margin.

As observed by Ben-David et al. [2], a matrix $M \in \{-1,1\}^{X\times Y}$ can always be realized by an arrangement of homogeneous half spaces with margin $\max(|X|^{-1/2}, |Y|^{-1/2})$: In the case $|X| \leqslant |Y|$ let, for $x \in X$, u_x be the canonical unit vector $e_x \in \mathbb{R}^X$, and let, for $y \in Y$, $v_y := |X|^{-1/2}(M_{x,y})_{x\in X} \in \mathbb{R}^X$. Then $\|u_x\| = \|v_y\| = 1$, $M_{x,y} = \operatorname{sign}\langle u_x, v_y\rangle$ and $|\langle u_x, v_y\rangle| = |X|^{-1/2}$ for all $x \in X$, $y \in Y$.

A concept class $(X, \mathscr{C})$ consists of a set X, called the instance space, and a set $\mathscr{C}$ of concepts, where any subset of X is called a concept. For a concept class $(X, \mathscr{C})$, a theorem by Vapnik [16,4, Theorem 4.16] gives an upper bound on the margin of any arrangement of half spaces that realizes $\mathrm{M}_{X,\mathscr{C}}$: There it is shown that if d is the VC-dimension of $(X, \mathscr{C})$, then at most the margin $d^{-1/2}$ is possible. It is always true that $d \leqslant |X|$, and if $d = |X|$ then the upper bound on the margin meets the lower bound from the trivial embedding given above. However, $d = |X|$ is a very special case: $d = |X|$ means that the set of concepts is the power set, $\mathscr{C} = \mathscr{P}(X)$.

Ben-David et al. [2] show that most concept classes (even of small VC-dimension) cannot be realized by an arrangement of half spaces with a margin that is much larger than the margin achieved by the trivial embedding. They use counting arguments that do not give an upper bound on the margin for explicit concept classes.

We show a result that implies that for some concrete concept classes (even if the VC-dimension is much smaller than $|X|$ and $|\mathscr{C}|$) the margin cannot be much larger than the margin achieved by the trivial embedding. In particular, it will follow from Theorem 3.1 that the trivial embedding gives the best possible margin if the rows or

the columns of $M_{X,\mathscr{C}}$ are orthogonal, or also if $M_{X,\mathscr{C}}$ has $|X|$ pairwise orthogonal columns.

Theorem 3.1. *If a matrix $M \in \{-1,1\}^{X\times Y}$ can be realized by an arrangement of homogeneous half spaces with margin γ, then*

$$\gamma \leqslant \frac{\|M\|}{\sqrt{|X|\,|Y|}}.$$

Proof. Let $u_x, v_y \in S^{k-1}$ be vectors with $\operatorname{sign}\langle u_x, v_y\rangle = M_{x,y}$ and $|\langle u_x, v_y\rangle| \geqslant \gamma$ for $x\in X$, $y\in Y$. From Lemma 2.1, it follows that $|Y|(|X|\gamma)^2 \leqslant |X|\,\|M\|^2$. □

For orthogonal matrices, in particular for Hadamard matrices, this shows that the trivial embedding gives the optimal margin:

Corollary 3.1. *The largest margin of any arrangement of homogeneous half spaces that realizes the concept class $(X,\mathscr{C})$ for which $M_{X,\mathscr{C}}$ is an Hadamard matrix H_n is $\gamma = 2^{-n/2}$.*

Proof. The trivial embedding has this margin, and Theorem 3.1 shows that this margin is optimal. □

Note that the VC-dimension d of the concept class of Corollary 3.1 is n. Thus, for this concept class the upper bound $d^{-1/2} = n^{-1/2}$ on the margin in terms of the VC-dimension is much weaker than our bound from Theorem 3.1.

4. A result from matrix theory

In this section, we show a result that was needed in the proof of Theorem 2.2. We start by defining some notation used in this section.

For any finite set $X \subseteq \mathbb{R}^k$ we consider the following positive semidefinite matrix:

$$M(X) := \sum_{x\in X} xx^{\top} \in \mathbb{R}^{k\times k}.$$

This matrix and its eigenvalues and eigenvectors can be used to measure in which directions the vectors $x\in X$ are pointing on the average. The range of the matrix $M(X)$ is $\operatorname{range}(M(X)) = \operatorname{span}(X)$. (This simple fact is, for example, shown in [8].) For any nonsingular linear transformation $A\in\mathbb{R}^{k\times k}$ we write $A(X) := \{Ax \mid x\in X\}$. We also write $N(X) := \{N(x) \mid x\in X\}$, where $N : \mathbb{R}^k\setminus\{0\} \to S^{k-1}$, $N(x) := \frac{x}{\|x\|}$, normalizes vectors. If $X\subseteq\mathbb{R}^k$, $k\leqslant|X|<\infty$, has the property that any subset of X with k elements is linearly independent, then so do the sets $A(X)$ for any nonsingular linear transformation $A\in\mathbb{R}^{k\times k}$ and the set $N(X)$. Furthermore, $|A(X)| = |N(X)| = |X|$. Repeated applications of nonsingular linear transformations and of normalizations N can be merged: Obviously $N(B(N(A(X)))) = N((B\circ A)(X))$ for any nonsingular linear transformations $A, B\in\mathbb{R}^{k\times k}$.

We want to prove that

Theorem 4.1. *Let $X\subseteq\mathbb{R}^k$, $|X|\geqslant k$, be a finite set such that all subsets of X with k elements are linearly independent. Then there is a nonsingular linear transformation*

$A \in \mathbb{R}^{k\times k}$ *such that*

$$\mathrm{M}(N(A(X))) = \sum_{x\in X} \frac{1}{\|Ax\|^2}(Ax)(Ax)^\top = \frac{|X|}{k} I_k.$$

Note that it is easy to find a nonsingular linear transformation $A \in \mathbb{R}^{k\times k}$ such that $\mathrm{M}(A(X))$ is "optimally balanced", i.e. such that $\mathrm{M}(A(X)) = I_k$. (How this can be done is shown below in the proof of Lemma 4.1.) However, this observation does not suffice to prove Theorem 4.1 because we are considering $\mathrm{M}(N(A(X)))$ and not $\mathrm{M}(A(X))$ there. But we can still use the linear transformation A with $\mathrm{M}(A(X)) = I_k$ to get a little closer to the solution we are looking for. This is formalized in Lemma 4.1. With this result and with a compactness argument (Lemma 4.2) we can finally prove Theorem 4.1.

Lemma 4.1. *Let $X \subseteq S^{k-1}$, $|X| \geqslant k$, be a finite set such that all subsets of X with k elements are linearly independent. Then either $\mathrm{M}(X) = \frac{|X|}{k} I_k$ or there is some nonsingular linear transformation $A \in \mathbb{R}^{k\times k}$ such that the smallest eigenvalue of*

$$\mathrm{M}(N(A(X))) = \sum_{x\in X} \frac{1}{\|Ax\|^2}(Ax)(Ax)^\top$$

is strictly larger than the smallest eigenvalue of $\mathrm{M}(X)$.

Proof. Let λ be the smallest eigenvalue of $\mathrm{M}(X)$ and m its multiplicity. The sum of the k (nonnegative) eigenvalues of $\mathrm{M}(X)$ is $\mathrm{trace}(\mathrm{M}(X)) = \sum_{x\in X} \|x\|^2 = |X|$. Thus, the smallest eigenvalue λ is at most $\frac{|X|}{k}$. Furthermore, if $\lambda = \frac{|X|}{k}$ then all eigenvalues of $\mathrm{M}(X)$ must be equal to $\frac{|X|}{k}$, i.e. $\mathrm{M}(X) = \frac{|X|}{k} I_k$. Thus, we can assume that $\lambda < \frac{|X|}{k}$. In this case, $m < k$ must also hold.

If $|X| = k$, we can simply map the elements of X to the canonical unit vectors with a nonsingular linear transformation A. Thus, we can assume that $|X| > k$.

We only show that there is a nonsingular linear transformation $A \in \mathbb{R}^{k\times k}$ such that the smallest eigenvalue of $\mathrm{M}(N(A(X)))$ is at least λ, and the multiplicity of the eigenvalue λ of $\mathrm{M}(N(A(X)))$ is strictly smaller than m. This suffices because we can repeat the procedure until the multiplicity of λ is zero.

We know that $\lambda \geqslant 0$ because $\mathrm{M}(X)$ is positive semidefinite. We even know that $\lambda > 0$ because there are k linearly independent elements in X which means that $\mathrm{M}(X)$ is nonsingular.

We can assume without loss of generality (using the spectral theorem for symmetric matrices) that

$$\mathrm{M}(X) = \mathrm{diag}(\lambda_1, \ldots, \lambda_k),$$

where $0 < \lambda = \lambda_1 = \cdots = \lambda_m < \lambda_{m+1} \leqslant \cdots \leqslant \lambda_k$. For the matrix $A := \mathrm{diag}(\lambda_1^{-1/2}, \ldots, \lambda_k^{-1/2}) \in \mathbb{R}^{k\times k}$ we have

$$\sum_{x\in X} (Ax)(Ax)^\top = A\mathrm{M}(X)A^\top = I_k. \tag{7}$$

Furthermore, for all $x \in X$:

$$\frac{1}{\|Ax\|^2} = \left(\sum_{i=1}^{k} \frac{x_i^2}{\lambda_i}\right)^{-1} \geqslant \lambda \tag{8}$$

because $\sum_{i=1}^{k} x_i^2 = \|x\|^2 = 1$ and $\lambda_1, \ldots, \lambda_k \geqslant \lambda$. Equality in (8) holds if and only if we have that $x_i = 0$ or $\lambda_i = \lambda$ for $i = 1, \ldots, k$, i.e. iff x is an eigenvector of $\mathrm{M}(X) = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ for the eigenvalue λ. Because of $1 \leqslant m < k$ this happens for at most m elements x of X, i.e. for at least $|X| - m$ elements of X we have strict inequality in (8). It is not hard to see that because of this the rank of the matrix

$$\mathrm{M}(N(A(X))) - \lambda I_k \overset{(7)}{=} \sum_{x \in X} \underbrace{\left(\frac{1}{\|Ax\|^2} - \lambda \right)}_{\geqslant 0^{(8)}} (Ax)(Ax)^{\top}, \tag{9}$$

is at least $\min(|X| - m, k)$. We can argue as follows: There is a subset Y of X of cardinality $\min(|X| - m, k)$ such that strict inequality in (8) holds for all $x \in Y$. The elements of Y are linearly independent because Y is a subset of X with at most k elements. It follows that the matrix on the right-hand side of (9) has rank at least $|Y| = \min(|X| - m, k)$, because for all $x \in Y$ the coefficients of the summands in (9) are strictly positive and $\mathrm{range}(\mathrm{M}(Z)) = \mathrm{span}(Z)$ holds for any set of vectors $Z \subseteq \mathbb{R}^k$. Thus, the eigenspace of $\mathrm{M}(N(A(X)))$ for λ has dimension at most

$$k - \min(|X| - m, k) \overset{|X| > k}{<} m.$$

We still have to show that the matrix $\mathrm{M}(N(A(X)))$ has no eigenvalues strictly smaller than λ. This is equivalent to showing that the matrix on the left-hand side of (9) is positive semidefinite. This is true because the right-hand side of (9) is a sum of positive semidefinite matrices. □

Note that the spectral norm induces a topology on the set of matrices $\mathbb{R}^{k \times k}$: We say that a sequence of matrices $A_1, A_2, \ldots \in \mathbb{R}^{k \times k}$ converges to the matrix $A \in \mathbb{R}^{k \times k}$ iff the spectral norms $\|A_l - A\|$ converge to zero as $l \to \infty$. A subset $\mathscr{A}$ of $\mathbb{R}^{k \times k}$ is called bounded if $\sup_{A \in \mathscr{A}} \|A\|$ is finite.

Lemma 4.2. *Let $X \subseteq \mathbb{R}^k$, $|X| \geqslant k$, be a finite set such that all subsets of X with k elements are linearly independent. Then for every $\varepsilon > 0$ the set of nonsingular matrices $A \in \mathbb{R}^{k \times k}$ with spectral norm $\|A\| = 1$, and for which all eigenvalues of $\mathrm{M}(N(A(X)))$ are at least $1 + \varepsilon$, is compact.*

Proof. Obviously this set of matrices is a bounded subset of $\mathbb{R}^{k \times k}$, and we have to show that it is closed in $\mathbb{R}^{k \times k}$. For this, let $A_1, A_2, \ldots$ be a sequence of elements from the set that converges to some $A \in \mathbb{R}^{k \times k}$. Clearly $\|A\| = 1$ holds because the spectral norm is continuous.

We only have to show that A is nonsingular: Once we know this, it follows that the matrices

$$\mathrm{M}(N(A_l(X))) = \sum_{x \in X} \frac{1}{\|A_l x\|^2} (A_l x)(A_l x)^{\top}$$

converge to $\mathrm{M}(N(A(X)))$ as $l \to \infty$. This implies that the eigenvalues of the matrix $\mathrm{M}(N(A(X)))$ are at least $1 + \varepsilon$, because they are the limits of the eigenvalues of the matrices $\mathrm{M}(N(A_l(X)))$ (see [9, Theorem 8.3.4]).

The idea of our proof that A is nonsingular is to show that if A would be singular then most vectors in $N(A_l(X))$ would get arbitrarily close to range(A) as $l \to \infty$. This means that not enough "weight" would remain in range$(A)^\perp$ for all eigenvalues of $\mathrm{M}(N(A_l(X)))$ to be larger than $1+\varepsilon$.

Let $n := \dim(\mathrm{null}(A))$. We also have that

$$\dim(\mathrm{range}(A)^\perp) = k - \dim(\mathrm{range}(A)) = \dim(\mathrm{null}(A)) = n.$$

Therefore, we can choose an orthonormal basis $d_1, \ldots, d_n$ of range$(A)^\perp$ consisting of n vectors. Because the smallest eigenvalue of $\mathrm{M}(N(A_l(X)))$ is at least $1+\varepsilon$, we know that

$$d_i^\top \mathrm{M}(N(A_l(X)))d_i \geqslant 1+\varepsilon$$

for $i = 1, \ldots, n$ (see (2)). If we sum over i we get that

$$n(1+\varepsilon) \leqslant \sum_{i=1}^{n} d_i^\top \mathrm{M}(N(A_l(X)))d_i = \sum_{x \in X} \underbrace{\sum_{i=1}^{n} \left\langle d_i, \frac{A_l x}{\|A_l x\|} \right\rangle^2}_{(*)}.$$

Because of Bessel's Inequality (3), each term $(*)$ is at most one. For $x \in X \backslash \mathrm{null}(A)$ we know even more: from $\|Ax\| > 0$ it follows that

$$\frac{A_l x}{\|A_l x\|} \overset{l \to \infty}{\to} \frac{Ax}{\|Ax\|} \in \mathrm{range}(A),$$

and because of $d_i \in \mathrm{range}(A)^\perp$, $(*)$ converges to zero for $l \to \infty$. It follows that

$$n(1+\varepsilon) \leqslant |X \cap \mathrm{null}(A)| \leqslant n. \tag{10}$$

For the second inequality in (10) note that by the assumption on X we know that

$$\dim(\mathrm{span}(X \cap \mathrm{null}(A))) = \min(|X \cap \mathrm{null}(A)|, k).$$

This is obviously upper bounded by $n = \dim(\mathrm{null}(A))$. Because of $\|A\| = 1$, we know that $k > n$, therefore we must have $|X \cap \mathrm{null}(A)| \leqslant n$.

Now we just have to note that (10) can only hold if $n = 0$, i.e. if A is nonsingular. □

Proof of Theorem 4.1. We start with a nonsingular linear transformation A that maps k arbitrarily chosen elements of X to the canonical unit vectors of $\mathbb{R}^k$. Then the smallest eigenvalue of $\mathrm{M}(N(A(X)))$ is at least 1. In the case $|X| = k$ we are already done.

Now assume that $|X| > k$. Lemma 4.1 says that if $\mathrm{M}(N(A(X))) = \frac{|X|}{k} I_k$ does not already hold we can modify A such that the smallest eigenvalue of $\mathrm{M}(N(A(X)))$ increases. Thus, we can find an A and an $\varepsilon > 0$ such that the smallest eigenvalue of $\mathrm{M}(N(A(X)))$ is $1+\varepsilon$. Obviously, we can assume that $\|A\| = 1$ (replace the matrix A by $\|A\|^{-1}A$).

By Lemma 4.2, the set of all nonsingular $A \in \mathbb{R}^{k \times k}$, $\|A\| = 1$, for which the smallest eigenvalue of $\mathrm{M}(N(A(X)))$ is at least $1+\varepsilon$, is compact. The smallest eigenvalue of $\mathrm{M}(N(A(X)))$ is a continuous function of $A \in \mathbb{R}^{k \times k}$. (The smallest eigenvalue of a positive semidefinite matrix is equal to the smallest singular value, and the singular values depend continuously on the matrix, see, e.g., Golub and Van Loan

[9, Theorem 8.3.4]) This means that there is a nonsingular $A \in \mathbb{R}^{k\times k}$, $\|A\| = 1$, for which the smallest eigenvalue of $\mathrm{M}(N(A(X)))$ is maximal.

For this A we must have $\mathrm{M}(N(A(X))) = \frac{|X|}{k} I_k$, because otherwise we could apply Lemma 4.1 like before and increase the smallest eigenvalue of $\mathrm{M}(N(A(X)))$. But this would contradict the fact that the smallest eigenvalue of $\mathrm{M}(N(A(X)))$ is already maximal. □

Acknowledgments

The author wants to thank Hans Ulrich Simon for a lot of helpful comments and for telling him about a nice idea how random projection and results from communication complexity theory can be used to prove upper bounds on margins of arrangements of half spaces. The anonymous referees made a number of helpful comments. The author is also grateful to Shai Ben-David, Niels Schmitt, Eike Kiltz, Ingo Wegener, Satyanarayana Lokam and Sanjoy Dasgupta for helpful discussions. We thank Dietrich Braess for simplifying the proof of Lemma 2.1. The author was supported by the Deutsche Forschungsgemeinschaft Grant SI 498/4-1 and by a Grant from the G.I.F., the German–Israeli Foundation for Scientific Research and Development.

References

[1] N. Alon, P. Frankl, V. Rödl, Geometrical realization of set systems and probabilistic communication complexity, Proceedings of the 26th Annual Symposium on Foundations of Computer Science, IEEE Computer Society, Portland, OR, 1985.

[2] S. Ben-David, N. Eiron, H.U. Simon, Limitations of learning via embeddings in Euclidean half-spaces (David P. Helmbold and Bob Williamson, Eds.), Proceedings of the 14th Annual Workshop on Computational Learning Theory, Springer, Berlin, Heidelberg, 2001.

[3] A. Blumer, A. Ehrenfeucht, D. Haussler, M.K. Warmuth, Learnability and the Vapnik–Chervonenkis dimension, J. ACM 100 (1989) 157–184.

[4] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, Cambridge, UK, 2000.

[5] J. Forster, A linear lower bound on the unbounded error probabilistic communication complexity (Francis M. Titsworth, Ed.), Proceedings of the 16th IEEE Annual Conference on Computational Complexity, IEEE Computer Society Press, Silver Spring, MD, 2001, pp. 100–106.

[6] J. Forster, M. Krause, S.V. Lokam, R. Mubarakzjanov, N. Schmitt, H.U. Simon, Relations between communication complexity, linear arrangements and computational complexity, Proceedings of the 21st Conference on Foundations of Software Technology and Theoretical Computer Science, Springer, Berlin, 2001, pp. 171–182.

[7] J. Forster, N. Schmitt, H.U. Simon, Estimating the optimal margins of embeddings in Euclidean half spaces, Proceedings of the 14th Annual Conference on Computational Learning Theory, Springer, Berlin, 2001, pp. 402–415.

[8] J. Forster, M.K. Warmuth, Relative loss bounds for temporal-difference learning, Proceedings of the Seventeenth International Conference on Machine Learning, Morgan Kaufmann, San Francisco, 2000, pp. 295–302.

[9] G.H. Golub, C.F. Van Loan, Matrix Computations, The Johns Hopkins University Press, Baltimore and London, 1991.

[10] R.A. Horn, C.R. Johnson, Matrix analysis, Cambridge University Press, Cambridge, UK, 1985.

[11] M.J. Kearns, U.V. Vazirani, An Introduction to Computational Learning Theory, MIT Press, Cambridge, MA, 1994.

[12] M. Krause, Geometric arguments yield better bounds for threshold circuits and distributed computing, Theoret. Comput. Sci. 156 (1996) 99–117.

[13] E. Kushilevitz, N. Nisan, Communication Complexity, Cambridge University Press, Cambridge, UK, 1997.

[14] R. Paturi, J. Simon, Probabilistic communication complexity, J. Comput. System Sci. 33 (1986) 106–123.
[15] W. Rudin, Real and Complex Analysis, McGraw-Hill, New York, 1974.
[16] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[17] V.N. Vapnik, A.Y. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, Theory Probab. Appl. 16 (1971) 264–280.